What about the Provenance of Media Archives: a Tale of Time-Contextualisation

Sam Coppens, Erik Mannens, Rik Van de Walle[1]

As the current multimedia assets of each institution –be it a content producer, an archive, a cultural institution, library, or museum– are being governed in a 'geographically distributed' way by some long-term preservation and dissemination platform, contextualization of the provenance information of each individual object is still often forgotten. Solutions for automatic discovery of the provenance information or ways for publishing provenance on the Web, beyond using Semantic web technologies, are rare and no real intertwined time-contextualisation of your assets nowadays exists. As such the project Archipel initiates the dissemination and digital long-term preservation of the cultural heritage in Flanders, Belgium, and researches the problems encountered with digital long-term preservation. In this project, we developed a platform that harvests data coming from various institutions (libraries, archival institutions, art museums, and the broadcasters), preserves the data for the long-term and disseminates the data as Linked Open Data (LOD) Dublin Core records. To guarantee the long-term preservation of the harvested content, our platform has the necessary processes in place to keep the information intact and interpretable, in line with the Open Archival Information System (OAIS) reference model for the long-term preservation of information.

These processes rely heavily on the provenance information of the harvested data, but at the same time produce also a lot of provenance information. This provenance information is modelled using a semantic implementation of the PREMIS 2.0 data dictionary, i.e., PREMIS OWL. Our developed platform generates many different versions of the harvested data, i.e., metadata and referenced multimedia files, via its preservation processes. These resources, their previous versions and their provenance information, relating the different versions, will be published on the Web as LOD. When preserving information for the long-term and publishing the information as LOD at the same time, different problems arise. First of all, we need to have persistent URIs for our resources, which will publish the information of a certain version of the resource. Another problem involves the enrichments that occur on the resources before publishing them as LOD. These enrichments will not always remain valid over time.

We need a way for preserving the temporality of these enrichments. The last problem being tackled by the Archipel project is the publication of the provenance information on the Web which will allow automatic discovery of the provenance information. To solve these problems, our developed platform is extended with the Memento datetime content negotiation. This datetime content negotiation will allow to select the appropriate version, called memento in the Memento framework, of the archived information and to publish it on a persistent URI. This datetime content negotiation will also solve the problem of preserving the temporality of the enrichments of the archived information. The different versions of the archived information are linked to each other via their provenance information. To publish the provenance information of each version on the Web, we extended the Memento framework to offer provenance links using a special Hypertext Transfer Protocol (HTTP) link header for automatic discovery of the provenance information.

In this paper/talk, we present how our digital long-term preservation platform is able to publish the provenance information on the Web. We introduce our semantic layered metadata model, which allows the archive to deal with the diversity of metadata records coming from diverse institutions and to track the provenance of the harvested data. We describe the distributed architecture of the archive and its processes. Finally, we explain the publication of the content and its provenance information using the Memento framework, extended to provide time-contextualisation of all of your assets.

[1] Ghent University – IBBT, Multimedia Lab, G. Crommenlaan 8, 9050 Ghent, Belgium